Brute force KNN model

In [0]:

```
```

In [1]:

```
# Load the Drive helper and mount
from google.colab import drive

# This will prompt for authorization.
drive.mount('/content/drive')
```

```
Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6
qk8qdgf4n4g3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%
b&scope=email%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdocs.test%20https%3A%2F%2Fwww.googleapis.
2Fauth%2Fdrive%20https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.photos.readonly%20https%3A%2F%2Fww
ogleapis.com%2Fauth%2Fpeopleapi.readonly&response_type=code

Enter your authorization code:
··········
Mounted at /content/drive
```

In [0]:

```
import warnings
warnings.filterwarnings("ignore")

import sqlite3
import pandas as pd
import numpy as np
import nltk
import string
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import roc_curve, auc
from nltk.stem.porter import PorterStemmer

import re
# Tutorial about Python regular expressions: https://pymotw.com/2/re/
import string
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer

from gensim.models import Word2Vec
from gensim.models import KeyedVectors
import pickle

# using the SQLite Table to read data.
con = sqlite3.connect('database.sqlite')

#filtering only positive and negative reviews i.e.
# not taking into consideration those reviews with Score=3
filtered_data = pd.read_sql_query(""" SELECT * FROM Reviews WHERE Score != 3 """, con)

# Give reviews with Score>3 a positive rating, and reviews with a score<3 a negative rating.
def partition(x):
    if x < 3:
        return 0
    return 1

#changing reviews with score less than 3 to be positive and vice-versa
```

```
#Changing reviews with Score less than 3 to be positive and vice-versa
actualScore = filtered_data['Score']
positiveNegative = actualScore.map(partition)
filtered_data['Score'] = positiveNegative
print(filtered_data.shape)
```

In [0]:

```
sorted_data=filtered_data.sort_values('ProductId', axis=0, ascending=True, inplace=False, kind='qui
cksort', na_position='last')
final=sorted_data.drop_duplicates(subset={"UserId","ProfileName","Time","Text"}, keep='first', inpl
ace=False)
final.shape
```

In [0]:

```
final.head(2)
```

In [0]:

```
final=final[final.HelpfulnessNumerator<=final.HelpfulnessDenominator]
#Before starting the next phase of preprocessing lets see the number of entries left
print(final.shape)

#How many positive and negative reviews are present in our dataset?
final['Score'].value_counts()
```

In [0]:

```
final.sort_values('Time',axis=0,ascending=True,inplace=True,kind='quicksort')
final.head(2)
```

In [0]:

```
stop = set(stopwords.words('english')) #set of stopwords
sno = nltk.stem.SnowballStemmer('english') #initialising the snowball stemmer

def cleanhtml(sentence): #function to clean the word of any html-tags
    cleanr = re.compile('<.*?>')
    cleantext = re.sub(cleanr, ' ', sentence)
    return cleantext
def cleanpunc(sentence): #function to clean the word of any punctuation or special characters
    cleaned = re.sub(r'[?|!|\'|"|#]',r'',sentence)
    cleaned = re.sub(r'[.|,|)|(|\|/]',r' ',cleaned)
    return  cleaned
```

In [0]:

```
#Code for implementing step-by-step the checks mentioned in the pre-processing phase
# this code takes a while to run as it needs to run on 500k sentences.
i=0
str1=' '
final_string=[]
all_positive_words=[] # store words from +ve reviews here
all_negative_words=[] # store words from -ve reviews here.
s=''
for sent in final['Text'].values:
    filtered_sentence=[]
    #print(sent);
    sent=cleanhtml(sent) # remove HTML tags
    for w in sent.split():
        for cleaned_words in cleanpunc(w).split():
            if((cleaned_words.isalpha()) & (len(cleaned_words)>2)):
                if(cleaned_words.lower() not in stop):
                    s=(sno.stem(cleaned_words.lower())).encode('utf8')
                    filtered_sentence.append(s)
                    if (final['Score'].values)[i] == 1:
                        all_positive_words.append(s) #list of all words used to describe positive r
eviews
                    if(final['Score'].values)[i] == 0:
                        all_negative_words.append(s) #list of all words used to describe negative r
eviews reviews
```

```
                else:
                    continue
            else:
                continue
    #print(filtered_sentence)
    str1 = b" ".join(filtered_sentence) #final string of cleaned words
    #print("*****************************************************************")

    final_string.append(str1)
    i+=1
```

In [0]:

```
final['CleanedText']=final_string #adding a column of CleanedText which displays the data after pr
e-processing of the review
final['CleanedText']=final['CleanedText'].str.decode("utf-8")
final.head(3)
```

In [0]:

```
import pickle
pickle.dump(final, open('drive/My Drive/Colab Notebooks/knn/final.p', 'wb'))
#final_sent = pickle.load(open('data.p','rb'))
final.shape
```

In [1]:

```
import pickle
final = pickle.load(open('drive/My Drive/Colab Notebooks/gbdt/final.p','rb'))


from sklearn.model_selection import train_test_split
##Sorting data according to Time in ascending order for Time Based Splitting
time_sorted_data = final.sort_values('Time', axis=0, ascending=True, inplace=False, kind='quicksort
', na_position='last')
final.head(2)
```

Out[1]:

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Sumn
138706	150524	0006641040	ACITT7DI6IDDL	shari zychinski	0	0	1	939340800	EV bo educati
138683	150501	0006641040	AJ46FKXOVC7NR	Nicholas A Mesiano	2	2	1	940809600	This w seri great to sp time

In [2]:

```
# sampling the data for ease
y = final['Score'].iloc[0:40000]
x = final['CleanedText'].iloc[0:40000]
x.shape,y.shape
```

Out[2]:

```
((40000,), (40000,))
```

BOW

In [3]:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

X_tra, X_tes, y_train, y_test = train_test_split(x,y,test_size=0.3,shuffle=False,random_state=0)

#Implementing BAG of words
bow = CountVectorizer()
X_tf =bow.fit_transform(X_tra)

# Standerdising the data
norm = StandardScaler(with_mean = False)
X_train = norm.fit_transform(X_tf)

# tfidf test
X_tfte = bow.transform(X_tes)

# Standerdising the data
X_test = norm.transform(X_tfte)
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/utils/validation.py:475: DataConversionWarning: Dat
a with input dtype int64 was converted to float64 by StandardScaler.
  warnings.warn(msg, DataConversionWarning)
```

In [7]:

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.cross_validation import cross_val_score
from sklearn import linear_model
from sklearn.metrics import make_scorer
from sklearn.metrics import f1_score

# 10 fold Cross Validation
myList = list(range(0,15))
neighbors = list(filter(lambda x: x % 2 != 0, myList))# Finding prime numbers

cv_scores = []

for k in neighbors:
    knn = KNeighborsClassifier(n_neighbors=k,algorithm='brute')
    scores = cross_val_score(knn, X_train,y_train, cv=10, scoring='f1_weighted')
    print('\nThe cross validation score for K = {} is {}.'.format(k,scores.mean()))
    cv_scores.append(scores.mean())

MSE = [1 - x for x in cv_scores]

optimal_k = neighbors[MSE.index(min(MSE))]
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/cross_validation.py:41: DeprecationWarning: This mo
dule was deprecated in version 0.18 in favor of the model_selection module into which all the refa
ctored classes and functions are moved. Also note that the interface of the new CV iterators are d
ifferent from that of this module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)
```

```
The cross validation score for K = 1 is 0.8409052274519677.

The cross validation score for K = 3 is 0.854205242845245.

The cross validation score for K = 5 is 0.8489215818144903.

The cross validation score for K = 7 is 0.8447328697023988.

The cross validation score for K = 9 is 0.8422545549630334.

The cross validation score for K = 11 is 0.8410442472820181.

The cross validation score for K = 13 is 0.8402215806876621.
```

In [8]:

```
print('\nThe optimal number of neighbors is %d.' % optimal_k)
```

```
The optimal number of neighbors is 3.
```

In [0]:

```
from sklearn.neighbors import KNeighborsClassifier
knn_optimal = KNeighborsClassifier(n_neighbors=3,algorithm='brute')

knn_optimal.fit(X_train,y_train)

pred = knn_optimal.predict(X_test)
```

In [11]:

```
from sklearn.metrics import f1_score,precision_score,recall_score
f1 = f1_score(y_test, pred, average='micro')
f1_per= f1*100

print('\nThe f1 score of the knn classifier of BOW for k = %d is %f%%' % (3, f1_per))
```

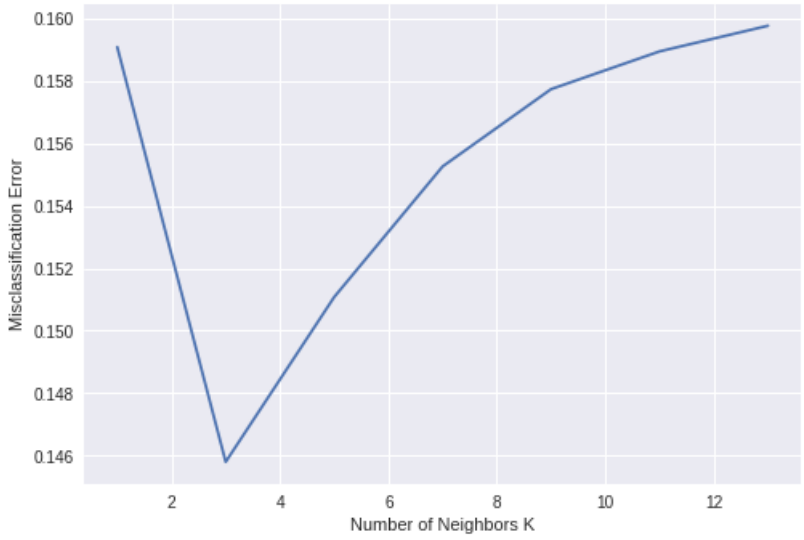
```
The f1 score of the knn classifier of BOW for k = 3 is 88.333333%
```

In [12]:

```
import matplotlib.pyplot as plt
import numpy as np
# plot misclassification error vs k
plt.plot(neighbors, MSE)


plt.xlabel('Number of Neighbors K')
plt.ylabel('Misclassification Error')
plt.show()

print("the misclassification error for each k value is : ", np.round(MSE,3))
```




```
the misclassification error for each k value is :  [0.159 0.146 0.151 0.155 0.158 0.159 0.16 ]
```

In [0]:

```

```

TFIDF

In [0]:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
X_tra, X_tes, y_train, y_test = train_test_split(x.values,y.values,test_size=0.3,shuffle=False)

#Implementing BAG of words
tfidf = TfidfVectorizer(ngram_range=(0,1),dtype=float)
X_tf =tfidf.fit_transform(X_tra)

# Standerdising the data
norm = StandardScaler(with_mean = False)
X_train = norm.fit_transform(X_tf)

# tfidf test
X_tfte = tfidf.transform(X_tes)

# Standerdising the data
X_test = norm.transform(X_tfte)
```

In [14]:

```
# 10 fold Cross Validation
myList = list(range(0,50))
neighbors = list(filter(lambda x: x % 2 != 0, myList))# Finding prime numbers

cv_scores = []

for k in neighbors:
    knn = KNeighborsClassifier(n_neighbors=k,algorithm='brute')
    scores = cross_val_score(knn, X_train,y_train, cv=10, scoring='f1_weighted')
    print('\nThe cross validation score for K = {} is {}.'.format(k,scores.mean()))
    cv_scores.append(scores.mean())

MSE = [1 - x for x in cv_scores]

optimal_k = neighbors[MSE.index(min(MSE))]
```

```
The cross validation score for K = 1 is 0.8419452942163795.

The cross validation score for K = 3 is 0.8450248777514654.

The cross validation score for K = 5 is 0.8414814566581768.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 7 is 0.8400123163639247.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 9 is 0.8393238274453758.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
```

```
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 11 is 0.8388895100916157.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 13 is 0.8388024646892724.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
```

```
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 15 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
```

```
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 17 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 19 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
```

```
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 21 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 23 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
```

```
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 25 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 27 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 29 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135: 
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 31 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 33 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 35 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 37 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 39 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 41 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 43 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
```

```
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 45 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 47 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
```

```
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

```
The cross validation score for K = 49 is 0.8387154192869289.
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: F-score is ill-defined and being set to 0.0 in labels with no predicted sa
mples.
  'precision', 'predicted', average, warn_for)
```

In [15]:

```
print('\nThe optimal number of neighbors is %d.' % optimal_k)
```

```
The optimal number of neighbors is 3.
```

In [0]:

```
knn_optimal = KNeighborsClassifier(n_neighbors=optimal_k,algorithm='brute')

knn_optimal.fit(X_train,y_train)

pred = knn_optimal.predict(X_test)
```

In [17]:

```
f1 = f1_score(y_test, pred, average='micro')
f1_per= f1*100


print('\nThe f1 score of the knn classifier of BOW for k = %d is %f%%' % (optimal_k, f1_per))
```

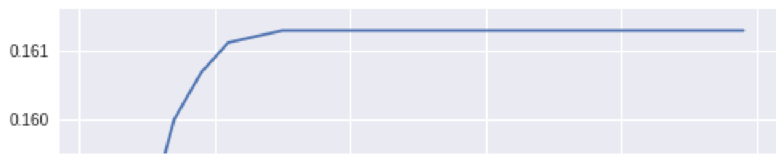
```
The f1 score of the knn classifier of BOW for k = 3 is 88.425000%
```

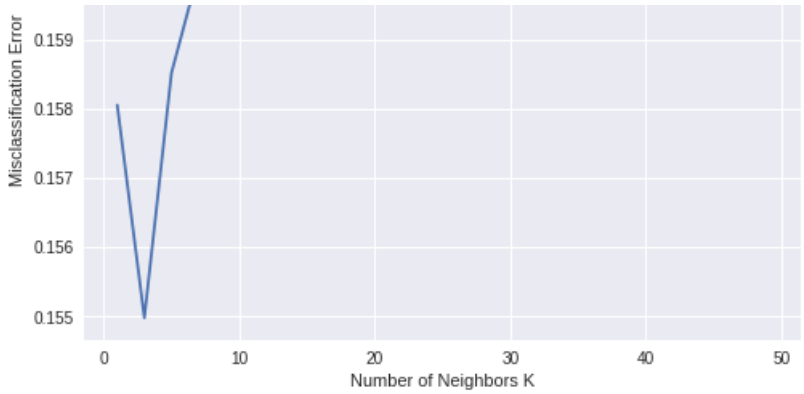
In [18]:

```
import matplotlib.pyplot as plt
import numpy as np
# plot misclassification error vs k
plt.plot(neighbors, MSE)


plt.xlabel('Number of Neighbors K')
plt.ylabel('Misclassification Error')
plt.show()

print("the misclassification error for each k value is : ", np.round(MSE,3))
```






```
the misclassification error for each k value is :  [0.158 0.155 0.159 0.16  0.161 0.161 0.161 0.161
0.161 0.161 0.161 0.161
 0.161 0.161 0.161 0.161 0.161 0.161 0.161 0.161 0.161 0.161 0.161 0.161
 0.161]
```

AVG-Word2vec

In [19]:

```
y = final['Score'].iloc[:50000]
x = final['CleanedText'].iloc[:50000]
x.shape,y.shape
```

Out[19]:

```
((50000,), (50000,))
```

In [20]:

```
%env JOBLIB_TEMP_FOLDER=/tmp
```

```
env: JOBLIB_TEMP_FOLDER=/tmp
```

In [0]:

```
X_tra, X_tes, y_train, y_test = train_test_split(x,y,test_size=0.3,shuffle=False)

sent_of_train=[]
for sent in X_tra:
    sent_of_train.append(sent.split())

sent_of_test=[]
for sent in X_tes:
    sent_of_test.append(sent.split())
```

In [22]:

```
!pip install gensim
```

```
Collecting gensim
  Downloading
https://files.pythonhosted.org/packages/27/a4/d10c0acc8528d838cda5eede0ee9c784caa598dbf40bd0911ff8c
7eb/gensim-3.6.0-cp36-cp36m-manylinux1_x86_64.whl (23.6MB)
    100% |████████████████████████████████| 23.6MB 1.9MB/s
Requirement already satisfied: six>=1.5.0 in /usr/local/lib/python3.6/dist-packages (from gensim)
(1.11.0)
Collecting smart-open>=1.2.1 (from gensim)
  Downloading
https://files.pythonhosted.org/packages/4b/1f/6f27e3682124de63ac97a0a5876da6186de6c19410feab66c1543
055/smart_open-1.7.1.tar.gz
Requirement already satisfied: numpy>=1.11.3 in /usr/local/lib/python3.6/dist-packages (from
gensim) (1.14.6)
Requirement already satisfied: scipy>=0.18.1 in /usr/local/lib/python3.6/dist-packages (from
gensim) (1.1.0)
Collecting boto>=2.32 (from smart_open>=1.2.1->gensim)
```

```
Collecting boto>=2.32 (from smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/23/10/c0b78c27298029e4454a472a1919bde20cb182dab1662cec7f2ca
523/boto-2.49.0-py2.py3-none-any.whl (1.4MB)
    100% |████████████████████████████████| 1.4MB 15.1MB/s
Collecting bz2file (from smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/61/39/122222b5e85cd41c391b68a99ee296584b2a2d1d233e7ee32b453
f2d/bz2file-0.98.tar.gz
Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-packages (from smart-
open>=1.2.1->gensim) (2.18.4)
Collecting boto3 (from smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/a0/d8/8b000ffeba218d47cd81fbd0bd0b2790742f81ffe116964a298be
8a4/boto3-1.9.50-py2.py3-none-any.whl (128kB)
    100% |████████████████████████████████| 133kB 26.9MB/s
Requirement already satisfied: idna<2.7,>=2.5 in /usr/local/lib/python3.6/dist-packages (from
requests->smart-open>=1.2.1->gensim) (2.6)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in /usr/local/lib/python3.6/dist-packages
(from requests->smart-open>=1.2.1->gensim) (3.0.4)
Requirement already satisfied: urllib3<1.23,>=1.21.1 in /usr/local/lib/python3.6/dist-packages
(from requests->smart-open>=1.2.1->gensim) (1.22)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-packages (from
requests->smart-open>=1.2.1->gensim) (2018.10.15)
Collecting jmespath<1.0.0,>=0.7.1 (from boto3->smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/b7/31/05c8d001f7f87f0f07289a5fc0fc3832e9a57f2dbd4d3b0fee70e
365/jmespath-0.9.3-py2.py3-none-any.whl
Collecting botocore<1.13.0,>=1.12.50 (from boto3->smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/82/c1/a26012b4dbca2e2ae06d7b24dc6b4378c0b0544d25f3e3f216612
033/botocore-1.12.50-py2.py3-none-any.whl (4.9MB)
    100% |████████████████████████████████| 4.9MB 6.8MB/s
Collecting s3transfer<0.2.0,>=0.1.10 (from boto3->smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/d7/14/2a0004d487464d120c9fb85313a75cd3d71a7506955be458eebfe
b1d/s3transfer-0.1.13-py2.py3-none-any.whl (59kB)
    100% |████████████████████████████████| 61kB 23.1MB/s
Requirement already satisfied: python-dateutil<3.0.0,>=2.1; python_version >= "2.7" in
/usr/local/lib/python3.6/dist-packages (from botocore<1.13.0,>=1.12.50->boto3->smart-open>=1.2.1->
gensim) (2.5.3)
Collecting docutils>=0.10 (from botocore<1.13.0,>=1.12.50->boto3->smart-open>=1.2.1->gensim)
  Downloading
https://files.pythonhosted.org/packages/36/fa/08e9e6e0e3cbd1d362c3bbee8d01d0aedb2155c4ac112b19ef3ca
d8d/docutils-0.14-py3-none-any.whl (543kB)
    100% |████████████████████████████████| 552kB 24.1MB/s
Building wheels for collected packages: smart-open, bz2file
  Running setup.py bdist_wheel for smart-open ... -   \   done
  Stored in directory:
/root/.cache/pip/wheels/23/00/44/e5b939f7a80c04e32297dbd6d96fa3065af89ecf57e2b5f89f
  Running setup.py bdist_wheel for bz2file ... -   done
  Stored in directory:
/root/.cache/pip/wheels/81/75/d6/e1317bf09bf1af5a30befc2a007869fa6e1f516b8f7c591cb9
Successfully built smart-open bz2file
Installing collected packages: boto, bz2file, jmespath, docutils, botocore, s3transfer, boto3, sma
rt-open, gensim
Successfully installed boto-2.49.0 boto3-1.9.50 botocore-1.12.50 bz2file-0.98 docutils-0.14 gensim
-3.6.0 jmespath-0.9.3 s3transfer-0.1.13 smart-open-1.7.1
```

In [0]:

```
pickle.dump(y_train, open('y_train.p', 'wb'))
pickle.dump(y_test, open('y_test.p', 'wb'))
```

In [23]:

```
#word to vector
from gensim.models import Word2Vec
w2v_model=Word2Vec(sent_of_train,min_count=3,size=200, workers=4)# words which occurs 3 times; 500
dimensions

w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 3 times ",len(w2v_words))
```

```
number of words that occured minimum 3 times  10559
```

In [24]:

```
# compute average word2vec for each review for X_train .
from tqdm import tqdm
import numpy as np



train_vectors = []
for sent in tqdm(sent_of_test):
    sent_vec = np.zeros(200)
    cnt_words =0;
    for word in sent:
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    train_vectors.append(sent_vec)


train_vectors1 = []
for sent in tqdm(sent_of_train):
    sent_vec = np.zeros(200)
    cnt_words =0;
    for word in sent:
        if word in w2v_words:
            vec = w2v_model.wv[word]
            sent_vec += vec
            cnt_words += 1
    if cnt_words != 0:
        sent_vec /= cnt_words
    train_vectors1.append(sent_vec)


len(train_vectors),len(train_vectors1)
```

```
100%|██████████| 15000/15000 [00:25<00:00, 588.50it/s]
100%|██████████| 35000/35000 [00:55<00:00, 629.57it/s]
```

Out[24]:

```
(15000, 35000)
```

In [25]:

```
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import TimeSeriesSplit

# Data-preprocessing: Standardizing the data
sc = StandardScaler(with_mean = False)
X_train3 = sc.fit_transform(train_vectors1)
X_test3 = sc.transform(train_vectors)


y_train.shape,X_train3.shape,X_test3.shape,y_test.shape
```

Out[25]:

```
((35000,), (35000, 200), (15000, 200), (15000,))
```

In [0]:

```
############################
X_train3 = pickle.load(open('drive/My Drive/Colab Notebooks/KNN/X_train3.p','rb'))
X_test3 = pickle.load(open('drive/My Drive/Colab Notebooks/KNN/X_test3.p','rb'))
```

In [0]:

In [26]:

```
X_train3.shape,X_test3.shape, y_train.shape, y_test.shape
```

Out[26]:

```
((35000, 200), (15000, 200), (35000,), (15000,))
```

In [27]:

```
#########################################################

from sklearn.neighbors import KNeighborsClassifier
from sklearn.cross_validation import cross_val_score
from sklearn.metrics import make_scorer
from sklearn.metrics import f1_score


# 10 fold Cross Validation
myList = list(range(0,15))
neighbors = list(filter(lambda x: x % 2 != 0, myList))

cv_scores = []

for k in neighbors:
    knn = KNeighborsClassifier(n_neighbors=k,algorithm='brute')
    scores = cross_val_score(knn,X_train3,y_train, cv=10, scoring='f1_weighted')
    print('\nThe cross validation score for K = {} is {}.'.format(k,scores.mean()))
    cv_scores.append(scores.mean())

MSE = [1 - x for x in cv_scores]

optimal_k = neighbors[MSE.index(min(MSE))]
```

```
The cross validation score for K = 1 is 0.85807776790818.

The cross validation score for K = 3 is 0.8729255886114174.

The cross validation score for K = 5 is 0.8755752538622973.

The cross validation score for K = 7 is 0.8749558733332183.

The cross validation score for K = 9 is 0.8739550089642041.

The cross validation score for K = 11 is 0.8731956450497907.

The cross validation score for K = 13 is 0.8722109996489932.
```

In [28]:

```
print('\nThe optimal number of neighbors is %d.' % optimal_k)
```

```
The optimal number of neighbors is 5.
```

In [0]:

```
from sklearn.neighbors import KNeighborsClassifier
knn_optimal = KNeighborsClassifier(n_neighbors=5,algorithm='brute')
```

In [0]:

```
knn_optimal.fit(X_train3,y_train)

pred = knn_optimal.predict(X_test3)
```

In [32]:

```
f1 = f1_score(y_test, pred, average='micro')
```

```
f1_sc =f1*100

print('\nThe f1 score of the knn classifier of BOW for k = %d is %f%%' % (5, f1_sc))
```

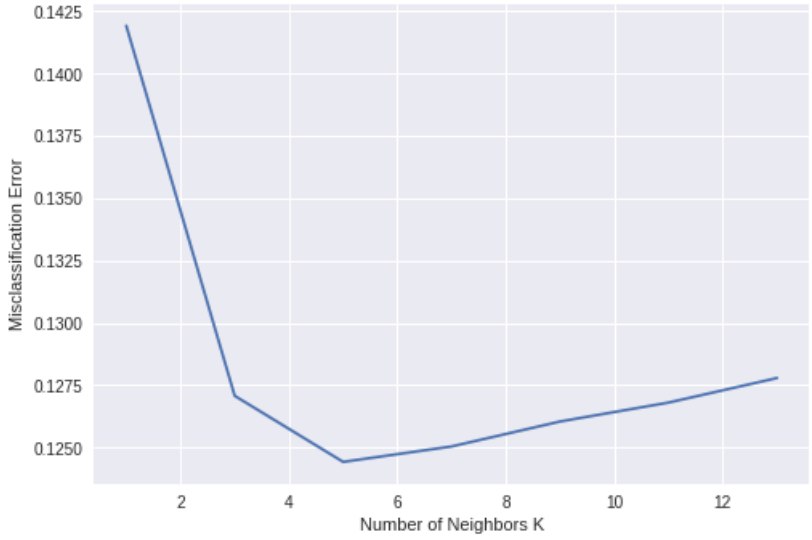
```
The f1 score of the knn classifier of BOW for k = 5 is 89.000000%
```

In [33]:

```
import matplotlib.pyplot as plt
import numpy as np
# plot misclassification error vs k
plt.plot(neighbors, MSE)


plt.xlabel('Number of Neighbors K')
plt.ylabel('Misclassification Error')
plt.show()

print("the misclassification error for each k value is : ", np.round(MSE,3))
```




```
the misclassification error for each k value is :  [0.142 0.127 0.124 0.125 0.126 0.127 0.128]
```

TFIDF-Word2vec

In [35]:

```
y = final['Score'].iloc[:50000]
x = final['CleanedText'].iloc[:50000]
x.shape,y.shape
```

Out[35]:

```
((50000,), (50000,))
```

In [0]:

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

X_tra, X_tes, y_train, y_test = train_test_split(x.values,y.values,test_size=0.3,random_state=0)
```

In [0]:

```
sent_of_train=[]
for sent in X_tra:
    sent_of_train.append(sent.split())
```

```
sent_of_test=[]
for sent in X_tes:
    sent_of_test.append(sent.split())
```

In [38]:

```
#word to vector
from gensim.models import Word2Vec
w2v_model=Word2Vec(sent_of_train,min_count=3,size=200, workers=4)# words which occurs 3 times; 500 dimensions

w2v_words = list(w2v_model.wv.vocab)
print("number of words that occured minimum 3 times ",len(w2v_words))
```

```
number of words that occured minimum 3 times  10497
```

In [0]:

```
m = TfidfVectorizer()
tf_idf_matrix = m.fit_transform(X_tra)
# we are converting a dictionary with word as a key, and the idf as a value
dictionary = dict(zip(m.get_feature_names(), list(m.idf_)))
```

In [40]:

```
from tqdm import tqdm
import numpy as np

# TF-IDF weighted Word2Vec
tfidf_feat = m.get_feature_names() # tfidf words/col-names
# final_tf_idf is the sparse matrix with row= sentence, col=word and cell_val = tfidf

tfidf_sent_vectors = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(sent_of_train): # for each review/sentence
    sent_vec = np.zeros(200) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors.append(sent_vec)
    row += 1
```

```
100%|██████████| 35000/35000 [01:02<00:00, 555.88it/s]
```

In [41]:

```
tfidf_sent_vectors1 = []; # the tfidf-w2v for each sentence/review is stored in this list
row=0;
for sent in tqdm(sent_of_test): # for each review/sentence
    sent_vec = np.zeros(200) # as word vectors are of zero length
    weight_sum =0; # num of words with a valid vector in the sentence/review
    for word in sent: # for each word in a review/sentence
        if word in w2v_words:
            vec = w2v_model.wv[word]
            tf_idf = dictionary[word]*(sent.count(word)/len(sent))
            sent_vec += (vec * tf_idf)
            weight_sum += tf_idf
    if weight_sum != 0:
        sent_vec /= weight_sum
    tfidf_sent_vectors1.append(sent_vec)
    row += 1
```

```
100%|██████████| 15000/15000 [00:30<00:00, 495.73it/s]
```

In [0]:

```
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import TimeSeriesSplit

# Data-preprocessing: Standardizing the data
sc = StandardScaler(with_mean = False)
X_train4 = sc.fit_transform(tfidf_sent_vectors)
X_test4 = sc.transform(tfidf_sent_vectors1)
```

In [43]:

```
X_train4.shape,X_test4.shape
```

Out[43]:

```
((35000, 200), (15000, 200))
```

In [0]:

```
import pickle

pickle.dump(X_train4, open('X_train4.p', 'wb'))
pickle.dump(X_test4, open('X_test4.p', 'wb'))
```

In [44]:

```
X_train4.shape,X_test4.shape, y_train.shape, y_test.shape
```

Out[44]:

```
((35000, 200), (15000, 200), (35000,), (15000,))
```

In [45]:

```
#########################################################

from sklearn.neighbors import KNeighborsClassifier
from sklearn.cross_validation import cross_val_score
from sklearn.metrics import make_scorer
from sklearn.metrics import f1_score


# 10 fold Cross Validation
myList = list(range(0,15))
neighbors = list(filter(lambda x: x % 2 != 0, myList))

cv_scores = []

for k in neighbors:
    knn = KNeighborsClassifier(n_neighbors=k,algorithm='brute')
    scores = cross_val_score(knn,X_train4,y_train, cv=10, scoring='f1_weighted')
    print('\nThe cross validation score for K = {} is {}.'.format(k,scores.mean()))
    cv_scores.append(scores.mean())

MSE = [1 - x for x in cv_scores]

optimal_k = neighbors[MSE.index(min(MSE))]
```

```
The cross validation score for K = 1 is 0.8479195225642234.

The cross validation score for K = 3 is 0.8637196918275543.

The cross validation score for K = 5 is 0.8662677218544383.

The cross validation score for K = 7 is 0.8657491941284302.

The cross validation score for K = 9 is 0.8651090093086712.

The cross validation score for K = 11 is 0.8630570690178787.
```

The cross validation score for K = 13 is 0.8610161430560955.

In [46]:

```
print('\nThe optimal number of neighbors is %d.' % optimal_k)
```

The optimal number of neighbors is 5.

In [0]:

```
x_t=np.array(X_train4)
y_t=np.array(y_train)
X_tes=np.array(X_test4)
```

In [0]:

```
knn_optimal = KNeighborsClassifier(n_neighbors=5,algorithm='brute')

knn_optimal.fit(X_train4,y_train)

pred = knn_optimal.predict(X_test4)
```

In [48]:

```
f1 = f1_score(y_test, pred, average='micro')

f1_sc =f1*100

print('\nThe f1 score of the knn classifier of BOW for k = %d is %f%%' % (5, f1_sc))
```

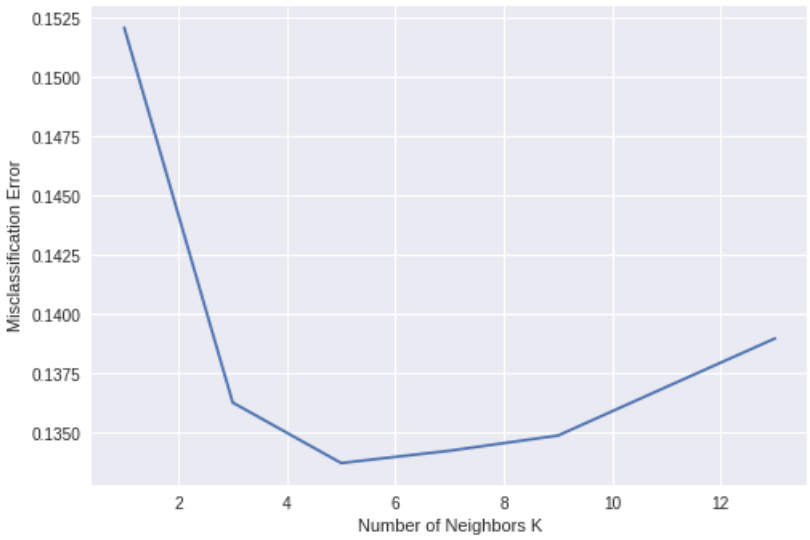
The f1 score of the knn classifier of BOW for k = 5 is 89.086667%

In [49]:

```
import matplotlib.pyplot as plt
import numpy as np
# plot misclassification error vs k
plt.plot(neighbors, MSE)


plt.xlabel('Number of Neighbors K')
plt.ylabel('Misclassification Error')
plt.show()

print("the misclassification error for each k value is : ", np.round(MSE,3))
```



the misclassification error for each k value is : [0.152 0.136 0.134 0.134 0.135 0.137 0.139]

==========

- F1 score of BOW at hyperparameter K = 3

```
83.88%
```

- F1 score of TFIDF at hyperparameter K = 3

```
88.45%
```

- F1 score of AVG W2V at hyperparameter K=5

```
89%
```

- F1 score of TFIDF W2V at hyperparameter k=5

```
89.09%
```

==========